Classification

Fall 2014

STAT 560

Classification/Discriminant Analysis

- For statisticians, these are synonyms
- Background: Two class (binary) version.
- Using "Training data" from Class 1 and Class 2
- Develop a "rule" for assigning new data to a Class
- Predicting whether a patient will develop breast cancer or remain healthy, given genetic information
- Predicting whether or not a user will like a new product, based on user covariates and a history of his/her previous ratings

Classification/Discriminant Analysis

There are a number of

- approaches
- philosophies
- thoughts

Often statistics vs EE-CS

Classification/Discriminant Analysis

EE-CS variations

- patten recognition
- artificial intelligence
- neural networks
- data mining
- machine learning

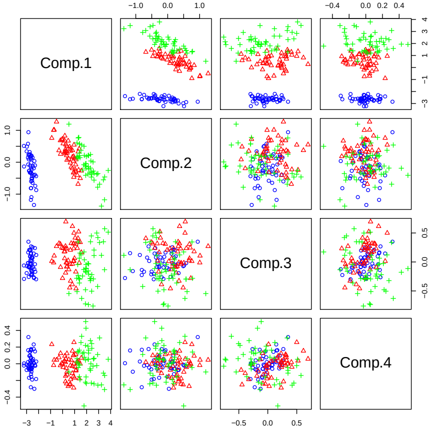
Classification/Discriminant Analysis

- Statistics viewpoint
  - Model "classes" with probability distributions
  - Use to study class differences and find rules
- EE-CS viewpoint
  - Data are set of numbers
  - Develop rules distinguish between these

Classification Basics

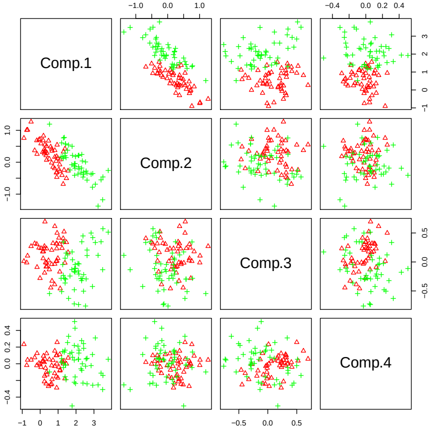
- Dat: $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ with the measurements of p (continuous) variables $\mathbf{x} \in \mathbb{R}^p$ and their class labels $y_i = 1, \ldots, K$ (categorical).
- Assume $\mathbf{x}|(y = k) \sim F_k$, for different distributions $\{F_i\}$.
- Binary class problems, $\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1} \sim F_1$ and $\mathbf{x}_{21}, \ldots, \mathbf{x}_{2n_2} \sim F_2$.
- Classification aims to classify a new observation, or several new observations into one of those classes.
- A *classifier* is a function $\phi : \mathcal{X} \to \{1, \ldots, K\}$.

Iris Data Example



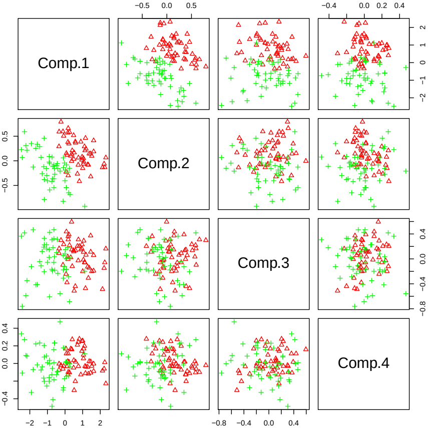
(Iris Setosa(b), Iris Versicolour (r), Iris Virginica(g))

Iris Data Example

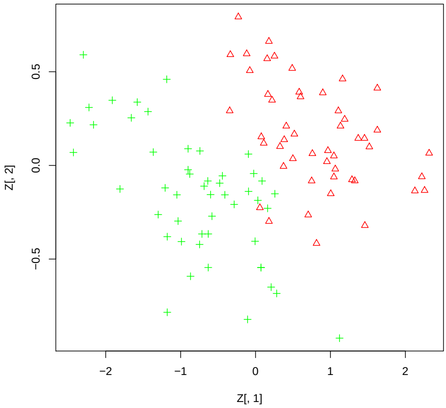


(Iris Versicolour (r), Iris Virginica(g))

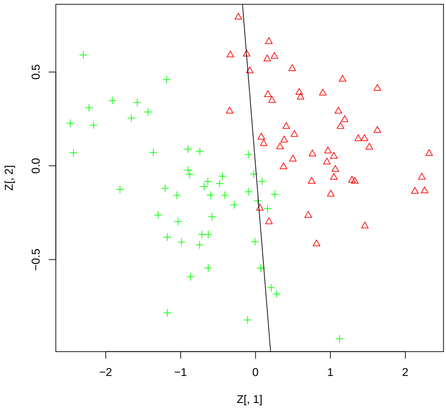
Iris Data Example



Iris Data Example



Iris Data Example

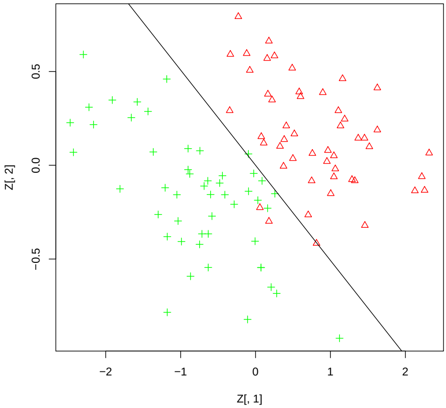


An example of classifier given by a linear hyperplain.

$\phi(\mathbf{z}) : \mathbb{R}^2 \to \{\text{versicolor(r)}, \text{virginica(g)}\}$, where

$$\phi(\mathbf{z}) = \begin{cases} \text{versicolor}, \mathbf{b}^\mathsf{T}\mathbf{z} > 0 \\ \text{virginica}, \mathbf{b}^\mathsf{T}\mathbf{z} \leq 0 \end{cases} \quad \mathbf{b} = \binom{-0.32}{1.58}$$

Iris Data Example



Another linear classifier $\phi(\mathbf{z}) : \mathbb{R}^2 \rightarrow \{\text{versicolor(r), virginica(g)}\}$,

where $\phi(\mathbf{z}) = \begin{cases} \text{versicolor}, \mathbf{b}^\mathsf{T}\mathbf{z} > 0 \\ \text{virginica}, \mathbf{b}^\mathsf{T}\mathbf{z} \leq 0 \end{cases}$ $\mathbf{b} = \binom{-5.28}{2.68}$

Example: Classifiers

The previous example on classifying Fisher's iris data is an example of linear classifier. A linear classifier $\phi(\mathbf{x})$ is a function of linear combinations of input vecot $\mathbf{x}$ and is of the form

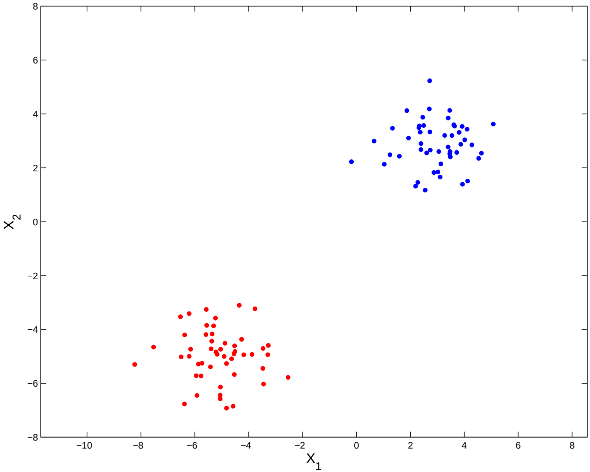
$$\phi(\mathbf{x}) = h(b_0 + \mathbf{b}^\top \mathbf{x})$$

In binary classification when $K = 2$, the linear classifier leads to the classification by a hyperplain. In general, a classifier can be of a higher order, for example, a quadratic classifier of the form

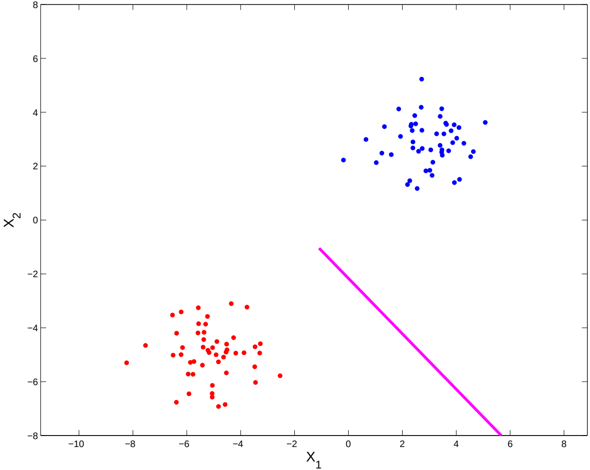
$$\phi(\mathbf{x}) = h(b_0 + \mathbf{v}^\top \mathbf{x} + \mathbf{x}^\top \mathbf{C}\mathbf{x})$$

or more general.

Example: Linear Classifier

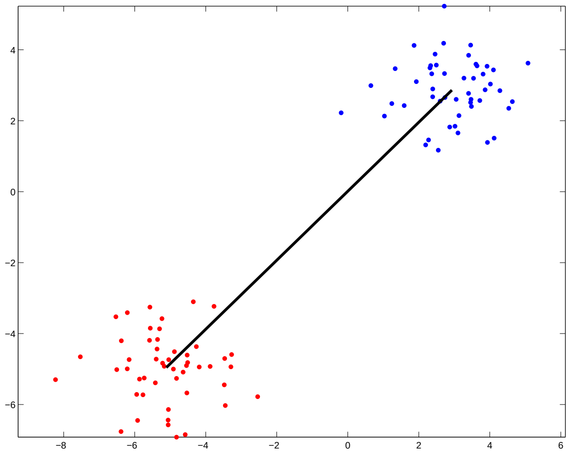


Example: Linear Classifier Decision Boundary

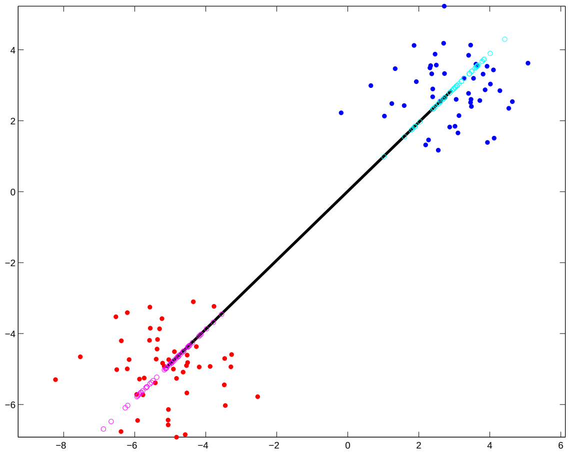


$$b_0 + x_1 b_1 + x_2 b_2 = 0$$

Example: Mean Difference



Example: Projection on MD

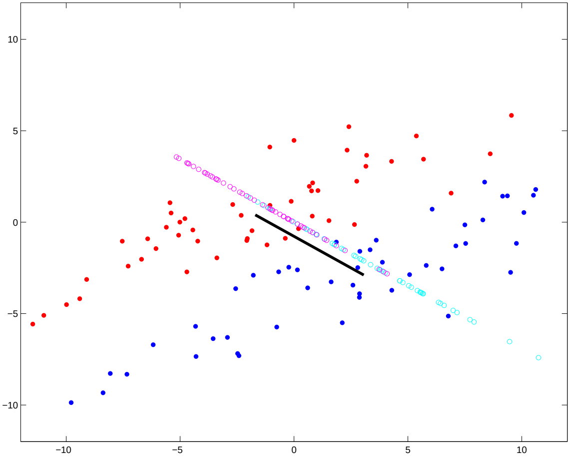


Linear Classifier

- Mean Difference (also known as Centroid Method):

$$\mathbf{b} \propto \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$$

Example 2: MD



Linear Classifiers

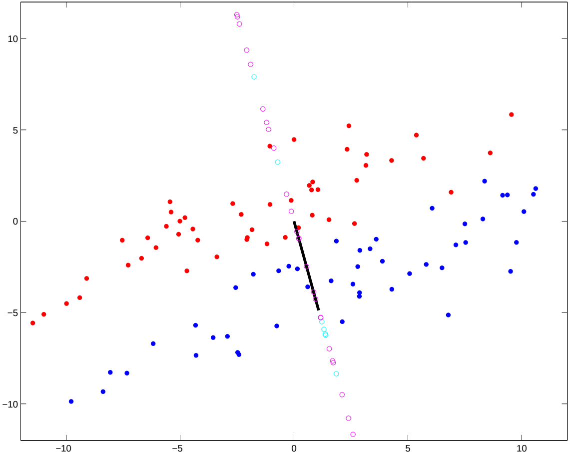
- Mean Difference:

$$\mathbf{b} \propto \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$$

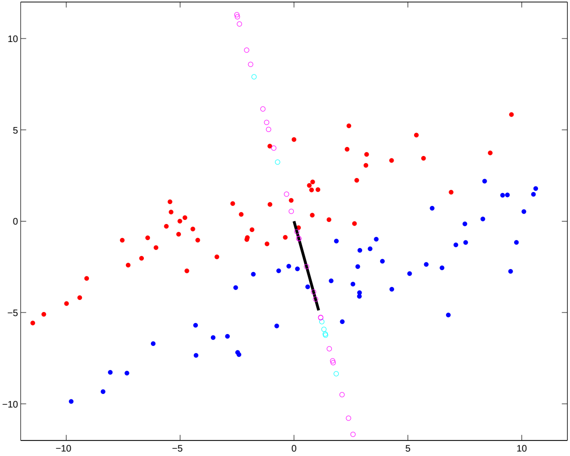
- Mean Difference on scaled data - Naive Bayes:

$$\mathbf{b} \propto \mathrm{Diag}(\mathbf{S})^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

Example 2: Naive Bayes



Example 2: Naive Bayes



why not adjusting for whole variance-covariance?

Linear Classifiers

- Mean Difference:

$$\mathbf{b} \propto \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$$

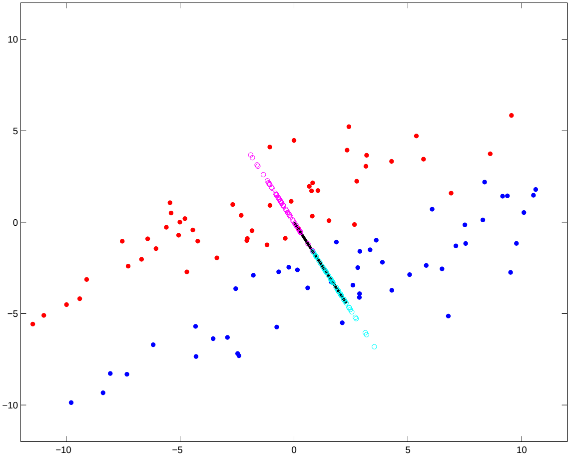
- Mean Difference on scaled data - Naive Bayes:

$$\mathbf{b} \propto \mathrm{Diag}(\mathbf{S})^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

- Linear Discrimnant Anlysis (a.k.a. Fisher's LDA):

$$\mathbf{b} \propto \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

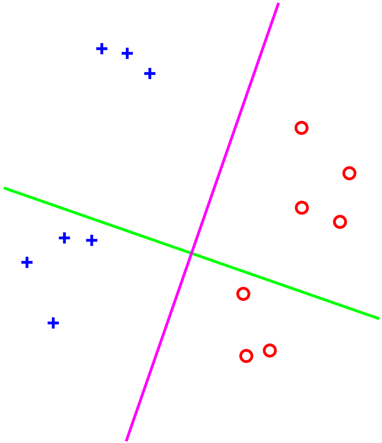
Example 2: LDA



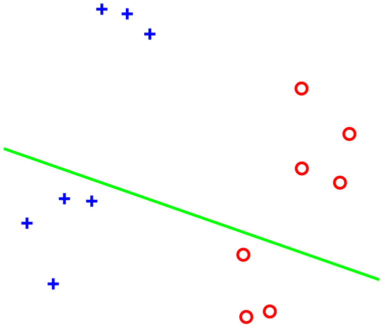
Major Assumption for LDA

- Class covariances are the same.

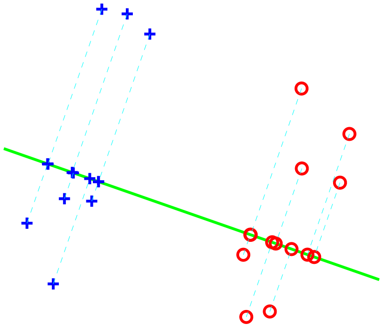
LDA



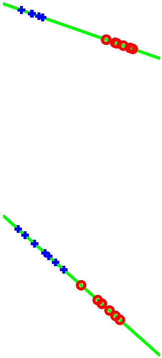
LDA



LDA

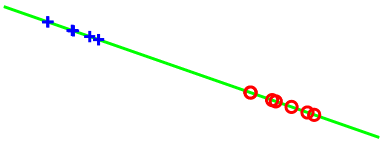


LDA



Which is BETTER?

LDA



$\bar{z}_1, \bar{z}_2$: projection means
s_1^2, s_2^2: scatter for projections.

LDA

- Fisher;s LDA direction $\mathbf{b}_{LDA}$ maximizes the ratio of the between-class projection variation to the within-class projection variation:

$$\mathbf{b}_{LDA} = \arg\max \frac{(\bar{z}_1 - \bar{z}_2)^2}{s_1^2 + s_2^2}.$$

Fisher Linear Discrimination

- For a direction vector $\mathbf{b} \in \mathbb{R}^p$
- Project the data onto the a line generated by $\mathbf{v}$:
  - Class 1: $\mathbf{b}^T\mathbf{x}_{11}, \ldots, \mathbf{b}^T\mathbf{x}_{1n_1}$
  - Class 2: $\mathbf{b}^T\mathbf{x}_{21}, \ldots, \mathbf{b}^T\mathbf{x}_{2n_2}$
- Then, the projection means are:
  - Class 1: $\bar{z}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbf{v}^T\mathbf{x}_{1i} = \mathbf{b}^T\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{x}_{1i}\right) = \mathbf{b}^T\bar{\mathbf{x}}_1$
  - Class 2: $\bar{z}_2 = \mathbf{b}^T\bar{\mathbf{x}}_2$.
- the projection scatters are:
  - Class 1:

$$
\begin{aligned}
s_1^2 &= \sum_{i=1}^{n_1}(\mathbf{b}^T\mathbf{x}_{1i} - \mathbf{b}^T\bar{\mathbf{x}}_1)^2 \\
&= \sum_{i=1}^{m}\mathbf{b}^T(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)^T\mathbf{b} \\
&= \mathbf{b}^T\mathbf{S}_1\mathbf{b}
\end{aligned}
$$

  - Class 2: $s_2^2 = \mathbf{b}^T\mathbf{S}_2\mathbf{b}$
- Thus, $\mathbf{b}_{LDA}$ maximizes $\frac{(\mathbf{b}^T\bar{\mathbf{x}}_1 - \mathbf{b}^T\bar{\mathbf{x}}_2)^2}{\mathbf{b}^T\mathbf{S}_1\mathbf{b} + \mathbf{b}^T\mathbf{S}_2\mathbf{b}} = \frac{\mathbf{b}^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{b}}{\mathbf{b}^T(\mathbf{S}_1 + \mathbf{S}_2)\mathbf{b}}$.

Linear Classifiers

- Mean Difference:

$$\mathbf{b} \propto \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$$

- Mean Difference on scaled data - Naive Bayes:

$$\mathbf{b} \propto \mathrm{Diag}(\mathbf{S})^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

- Fisher's LDA:

$$\mathbf{b} \propto \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

So far, these classifiers are introduced without any distributional assumption of the samples.

Bayesian Decision Theory

- Prior: $p(y)$, distribution of class lables
- Likelihood: $\mathbf{x}|(Y = k) \sim F_k$
- After seing the feature $\mathbf{x}$, what would be your action in assigining the class lable to this sample?

Bayesian Decision Theory

- Prior: $p(y)$, distribution of class lables
- Likelihood: $\mathbf{x}|(Y = k) \sim F_k$
- After seing the feature $\mathbf{x}$, what would be your action in assigining the class lable to this sample? Bayes rule assign the class which provides highest posterior probability:

$$\phi(\mathbf{x}) = \arg \max_{k=1,\ldots,K} P(Y = k|\mathbf{x})$$

- How to compute the posteior probability?

Bayesian Decision Theory

- Prior: $p(y)$, distribution of class lables
- Likelihood: $\mathbf{x}|(Y = k) \sim F_k$
- After seing the feature $\mathbf{x}$, what would be your action in assigining the class lable to this sample? Bayes rule assign the class which provides highest posterior probability:

$$\phi(\mathbf{x}) = \arg \max_{k=1,\dots,K} P(Y = k|\mathbf{x})$$

- How to compute the posteior probability? Bayes Theorem!

Bayesian Decision Theory: 1-d example

- Factory Production Line: I (x%) , II (100-x%)
- $X|(Y = I) \sim N(\mu_1, \sigma_1^2), X|(Y = II) \sim N(\mu_2, \sigma_2^2)$.
- After seing the feature **x**, what would be your action in assigining the class lables to this sample?
  Need Example here

Bayesian Decision Theory: 1-d example

- Posteir probability:

$$P(Y = i|\mathbf{x}) = \frac{f_i(\mathbf{x})\pi_i}{f_1(\mathbf{x})\pi_1 + f_2(\mathbf{x})\pi_2}$$

  - In English, *posterior* $= \frac{\textit{likelihood} \times \textit{prior}}{\text{marginal of features}}$
- Choose Production Line I if and only if

$$f_1(x)\pi_1 > f_2(x)\pi_2$$

Bayesian Decision Theory: 1-d example

- Case I: $\pi_1 = \pi_2, \sigma_1 = \sigma_2$ choose I if

$$|x - \mu_1| < |x - \mu_2|$$

- Case II: $\pi_1 = \pi_2, \sigma_1 \neq \sigma_2$ choose I if

$$(\frac{x - \mu_1}{\sigma_1^2})^2 < (\frac{x - \mu_2}{\sigma_2^2})^2 + 2\log(\sigma_2/\sigma_1)$$

- Case III: $\pi_1 \neq \pi_2, \sigma_1 = \sigma_2$ choose I if

$$(\frac{x - \mu_1}{\sigma_1^2})^2 < (\frac{x - \mu_2}{\sigma_2^2})^2 + 2\log(\pi_1/\pi_2)$$

Bayesian Decision Theory

- Assume that there are K different populations with $P(Y = i) = \pi_i$ and $\sum_{i=1}^{K} \pi_i = 1$.
- For each population, assume

$$\mathbf{x}|(Y = i) \sim f_i(\mathbf{x}).$$

- After seeing the features from the mixture of the K populations, Bayes theorem provides the posterior probability that the observed $\mathbf{x}$ was from the population i:

$$P(Y = i|\mathbf{x}) = \frac{f_i(\mathbf{x})\pi_i}{f_1(\mathbf{x})\pi_1 + \ldots f_K(\mathbf{x})\pi_K}$$

- Assign the class label which gives the highest posterior probaiblity.

Bayesian Decision Theory- Normal Density

- Now assume multivariate normal distributions: $\mathbf{x}|(Y = i) \sim N_p(\mu_i, \Sigma_i)$.
- Marginal distribution of features: $\mathbf{X} \sim f(\mathbf{x})$, a mixture of multivariate normals.
- Then, Bayes rule:

$$\begin{aligned}\arg\max P(Y = i|\mathbf{x}) = &= \arg\max \frac{f_i(\mathbf{x})\pi_i}{f_1(\mathbf{x})\pi_1 + \ldots f_K(\mathbf{x})\pi_K} \\ &= \arg\max f_i(\mathbf{x})\pi_i(\mathbf{x})\end{aligned}$$

where

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma_i^{-1}(\mathbf{x} - \mu_i)\right)$$

Bayesian Decision Theory- Normal Density

- One can write the Bayes rule in the following format:

$$\phi(\mathbf{x}) = \arg\min\{\log|\Sigma_i| + (\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i) - 2\log(\pi_i)\}$$

  - Special cases under common covariance assumption: $\Sigma_i = \Sigma$ and $\pi_i = 1/K$.
  - Bayes rule compares Mahalanobis distance

$$d_{MH}^2(\mathbf{x}, \mu_i) = (\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i)$$

    and assigns $\mathbf{x}$ to the nearest class based on this distance measure.
    Need Example Here!!

Bayesian Decision Theory- Normal Density

- One can write the Bayes rule in the following format:

$$\phi(\mathbf{x}) = \arg\min(b_{0i} + \mathbf{b}_i^\mathsf{T}\mathbf{x} + \mathbf{x}^\mathsf{T}\mathbf{C}_i\mathbf{x}), \text{ where}$$

$$\begin{aligned} b_{0i} &= \frac{1}{2}\mu_i^\mathsf{T}\Sigma_i^{-1}\mu_i + \frac{1}{2}\log|\Sigma_i| - \log(\pi_i), \\ \mathbf{b}_i &= -\Sigma_i^{-1}\mu_i \\ \mathbf{C}_i &= \frac{1}{2}\Sigma_i^{-1}. \end{aligned}$$

  - Quadratic Discriminant Analysis: Bayes rule classifier under unequal covariance
  - Linear Discriminant Analysis: Bayes rule classifier under equal covariance

$$\phi(\mathbf{x}) = \arg\min(b_{0i} + \mathbf{b}_i^\mathsf{T}\mathbf{x}).$$

Two class- Quadratic Discriminant Analysis

- The Bayes rule assigns $\mathbf{x}$ into class 1 if

$$b_{01} + \mathbf{b}_1^\top \mathbf{x} + \mathbf{x}^\top \mathbf{C}_1 \mathbf{x} \leq b_{02} + \mathbf{b}_2^\top \mathbf{x} + \mathbf{x}^\top \mathbf{C}_2 \mathbf{x}$$

Example Here

Two class - Linear Discriminant Analysis

- Assume the equal covariance $\Sigma_1 = \Sigma_2$.
- Then, the Bayes rule assigns $\mathbf{x}$ into class 1 if

$$
\begin{aligned}
&\underbrace{(\mu_2 - \mu_1)^\intercal \Sigma^{-1}}(\mathbf{x} - \frac{\mu_1 + \mu_2}{2}) < \log(\pi_1/\pi_2) \\
\Longleftrightarrow\ &\mathbf{b}^\intercal(\mathbf{x} - \mu) < \log(\pi_1/\pi_2),\ \text{where } \mathbf{b} = \Sigma^{-1}(\mu_2 - \mu_1).
\end{aligned}
$$

  - Special case: If $\Sigma = \mathbf{I}$, then $\mathbf{b} = (\mu_2 - \mu_1)$.
  - Special case: If $\Sigma =$ Diagonal Matrix , then $\mathbf{b} = \text{Diag}(\Sigma)^{-1}(\mu_2 - \mu_1)$.

Fisher's LDA

- LDA is often referred to as Fisher's LDA.
- His original work does not involve any distributional assumption, and develops LDA in a fashion similar to a geometric understanding of PCA.
- LDA direction seeks to maximizes the between group separation relative to the within-group variance of the projected scores.

Sample Bayes Rule Classifier

- In practice, we do not know the parameters (μ_i, Σ_i, π_i). Given observations

  $$\{(\mathbf{x}_{ij}, i), i = 1, \ldots, K, j = 1, \ldots, n_i, n = \sum_{i=1}^{K} n_i\},$$

  a sample version of the classifier is obtained by replacing parameters by their estimates, $\hat{\mu}_i = \bar{\mathbf{x}}_i, \hat{\Sigma}_i = \mathbf{S}_i$, and $\hat{\Sigma} = \mathbf{S}_P$ (pooled sample covariance matrix). Also, use $\hat{\pi}_i = n_i/n$.

Binary LDA and QDA classifiers

- QDA:

$$\phi(\mathbf{x}) = \arg\min_{i=1,2}(b_{0i} + \mathbf{b}_i^\top \mathbf{x} + \mathbf{x}^\top \mathbf{C}_i \mathbf{x})$$

- LDA: assign $\mathbf{x}$ into class 1 if

$$\mathbf{b}^\top(\mathbf{x} - \frac{\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2}{2}) < \log(\frac{n_1}{n_2})$$

where $\mathbf{b} = \mathbf{S}_P^{-1}(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1), \mathbf{S}_P = \frac{1}{n-2}((n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{X}_2)$

LDA vs QDA Examples